SCIENTIFIC REPORTS

OPEN

Wx: a neural network-based feature selection algorithm for transcriptomic data

Received: 2 October 2018
Accepted: 9 July 2019
Published online: 19 July 2019

Sungsoo Park[1], Bonggun Shin[1,3], Won Sang Shim[1], Yoonjung Choi[1], Kilsoo Kang[1] & Keunsoo Kang[2]

Next-generation sequencing (NGS), which allows the simultaneous sequencing of billions of DNA fragments simultaneously, has revolutionized how we study genomics and molecular biology by generating genome-wide molecular maps of molecules of interest. However, the amount of information produced by NGS has made it difficult for researchers to choose the optimal set of genes. We have sought to resolve this issue by developing a neural network-based feature (gene) selection algorithm called Wx. The Wx algorithm ranks genes based on the discriminative index (DI) score that represents the classification power for distinguishing given groups. With a gene list ranked by DI score, researchers can institutively select the optimal set of genes from the highest-ranking ones. We applied the Wx algorithm to a TCGA pan-cancer gene-expression cohort to identify an optimal set of gene-expression biomarker candidates that can distinguish cancer samples from normal samples for 12 different types of cancer. The 14 gene-expression biomarker candidates identified by Wx were comparable to or outperformed previously reported universal gene expression biomarkers, highlighting the usefulness of the Wx algorithm for next-generation sequencing data. Thus, we anticipate that the Wx algorithm can complement current state-of-the-art analytical applications for the identification of biomarker candidates as an alternative method. The stand-alone and web versions of the Wx algorithm are available at https://github.com/deargen/DearWXpub and https://wx.deargendev.me/, respectively.

Advances in science and technology often lead to paradigm shifts. In biology and biomedical fields, high-throughput screening (HTS) techniques such as microarray and next-generation sequencing (NGS) have changed how we identify measurable biological indicators (called biomarkers) for various diseases. For example, to identify biomarkers, which is how we to predict the onset or prognosis of various diseases, the conventional approach is mostly based on the manual selection of genes or particular loci on the genome with limited information from the literature. Then, experimental validation is required to confirm the biomarker selection. In this typical process, the initial selection of biomarkers is the most important and critical step.

Several sets of gene expression biomarkers have been developed and used to predict early diagnoses or to classify different sub-types of given diseases in clinics; for example, PAM50[1] has been successfully used to classify subtypes of breast cancer[2]. Recently, the HTS methodology has accelerated the process of identifying biomarkers, since this approach is capable of quantifying a whole set of molecules of interest accurately and simultaneously. For example, gene expression profiling based on the NGS technique (called RNA-seq) can accurately quantify the expression levels of whole genes in a given cell population. With a full list of genes (up to 190,000 transcripts in the human genome; https://www.gencodegenes.org/), researchers can narrow down biomarker candidates via downstream analyses such as unsupervised clustering[3], gene ontology (GO) analysis[4], regression analysis[5], and/or differentially expression gene (DEG) analysis[6,7]. Among these approaches, DEG analysis, which provides a list of genes (DEGs) that show significantly altered expressions between two or more groups with a statistical cutoff (adjusted p value) of 0.05, is widely used for the identification of biomarker candidates. However, the number of DEGs depends on the number of samples and the samples' characteristics. As the number of samples has increased due to the reduced sequencing cost, the number of DEGs has tended to increase to several thousand[8]. Therefore, it is difficult for researchers to choose the optimal combination of genes (biomarker candidates) from

[1]Deargen Inc., Daejeon, Republic of Korea. [2]Department of Microbiology, College of Natural Sciences, Dankook University, Cheonan, 31116, Republic of Korea. [3]Department of Computer Science, Emory University, Atlanta, GA, 30322, USA. Sungsoo Park and Bonggun Shin contributed equally. Correspondence and requests for materials should be addressed to K.K. (email: Kangk1204@dankook.ac.kr)

Type ID	Full name	# of cancer samples	# of normal samples	# of total samples	Ratio (cancer/total)
BLCA	Bladder urothelial carcinoma	408	19	427	0.95
BRCA	Breast invasive carcinoma	1101	113	1214	0.90
COAD	Colon adenocarcinoma	286	41	327	0.87
HNSC	Head and neck squamous cell carcinoma	522	44	566	0.92
KICH	Kidney chromophobe	65	25	90	0.72
KIRC	Kidney renal clear cell carcinoma	534	72	606	0.88
KIRP	Kidney renal papillary cell carcinoma	291	32	323	0.90
LIHC	Liver hepatocellular carcinoma	374	50	424	0.88
LUAD	Lung adenocarcinoma	517	59	576	0.89
LUSC	Lung squamous cell carcinoma	502	51	553	0.90
PRAD	Prostate adenocarcinoma	497	52	549	0.90
THCA	Thyroid carcinoma	512	59	571	0.89

Table 1. The number of cancer and normal samples used in this study.

the large number of DEGs using current approaches. Recently, several algorithms, which select features from high-dimensional NGS data, were reported[9,10]. Similar to these, we have sought to resolve the feature selection issue by developing a novel neural network-based feature (gene) selection algorithm called Wx. The Wx algorithm ranks genes based on their discriminative index (*DI*) score, which represents the classification power for distinguishing given groups. With a gene list ranked by *DI* score, researchers can institutively select an optimal set of genes from the highest-ranking ones. We tested the algorithm's usefulness by attempting to identify universal gene-expression cancer biomarker candidates that could potentially distinguish various types of cancer from normal samples in the pan-cancer data set of the cancer genome atlas (TCGA) project. The pan-cancer project was established to gain biological insights by defining commonalties and differences across cancer types and their organs of origin[11]. In addition to the pan-cancer RNA-seq data, three different cancer RNA-seq data from gene expression omnibus (GEO) were used to evaluate the performance of the identified biomarker candidates. Our algorithm successfully identified 14 key genes as a conceptual set of universal biomarkers, accurately distinguishing 12 types of cancer from normal tissue samples. The 14-gene signature was comparable to or outperformed previously reported universal gene expression biomarkers[12,13] in terms of classification accuracy. Further validation of the identified gene signature with three independent studies confirmed that the 14-gene signature identified by the Wx algorithm accurately classified cancer samples from normal samples compared to other methods[14,15]. Accordingly, we expect that the Wx algorithm can complement differentially expressed gene (DEG) analysis as an alternative method for the identification of biomarker candidates.

Results

We applied our Wx method, which is based on the Discriminative Index (*DI*) algorithm (see Methods), into a pan-cancer cohort from TCGA RNA-seq data consisting of 12 different types of cancer and normal (control) samples (Table 1). For this, a special case of the *DI*-based feature selection algorithm was constructed with only two labels (normal and cancer, $K = 2$). This analysis was intended to identify potential cross-cancer gene signatures (biomarkers) similar to a previous study[12]; we defined the identified biomarkers as universal gene-expression cancer biomarkers (UGCBs) for the pan-cancer cohort. Additional independent RNA-seq data from melanoma (GSE72056)[15], lung adenocarcinoma (GSE40419)[16], and head and neck squamous cell carcinoma single cells (GSE103322)[17] were used to evaluate identified UGCBs. The classification performance of the UGCBs identified by each approach was assessed by means of leave-one-out cross validation (LOOCV).

Identification of universal gene-expression cancer biomarkers. We identified the UGCBs distinguishing cancer samples from normal samples by applying the Wx algorithm to a pan-cancer cohort containing 6,226 total (5,609 tumor and 617 control) samples in 12 different types of cancers (Table 1). The samples in each cancer and their corresponding control group were randomly divided into two sets, a training set and validation set, which were used for feature selection and validation purposes, respectively. Because the Wx algorithm was based on a neural network method that trains the weights of network in the training set, the trained weight was highly dependent on the random values assigned to the initial value. Therefore, we avoided this irregularity by iterating the Wx algorithm 10,000 times and the highest genes (features) ranked by the average value of the *DI* score were selected as UGCBs. The entire list of genes with the averaged *DI* scores can be found in Table S1.

Comparison of UGCBs. We first determined how many genes from the gene list indexed by the *DI* score were required to maximize the average accuracy. For this, each set containing the top genes (1 to 1,000) was constructed to evaluate the average accuracy of cancer and normal sample classifications in the training set. Approximately the top 100 genes showed the highest average accuracy and no further increase in average accuracy was observed when more genes were added (Fig. 1).

Next, we selected the top 14 UGCBs (or top seven UGCBs) to compare the UGCBs reported in previous studies (Table 2). Interestingly, none of our UGCBs overlapped with those identified by Peng *et al.*[12] or Martinez-Ledesma *et al.*[13] (Table 2). Given that there were no common genes between independent studies, we wondered which sets of UGCBs would be the best in terms of classification accuracy. For this, the LOOCV method, which estimates the generalization performance of a given model trained on n – 1 samples and validates

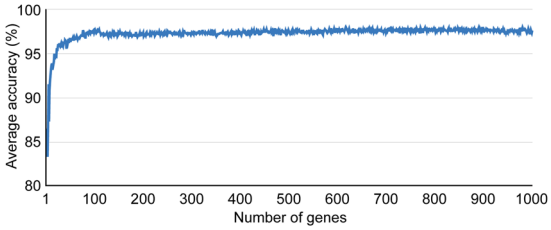

Figure 1. Classification accuracy according to given number of genes. The x-axis indicates the number of top genes (sorted in descending order by the *DI* values) used for the calculation and the y-axis represents the average accuracy.

Cancer type	Wx (this study)	Peng *et al.*	Emmanual *et al.*
BLCA	*EEF1A1*, *FN1*, *GAPDH*, *SFTPC*, *AHNAK*, *KLK3*, *UMOD*, *CTSB*, *COL1A1*, *GPX3*, *GNAS*, *ATP1A1*, *SFTPB*, *ACTB*	*KIF4A*, *NUSAP1*, *HJURP*, *NEK2*, *FANCI*, *DTL*, *UHRF1*, *FEN1*, *IQGAP3*, *KIF20A*, *TRIM59*, *CENPL*, *C16ORF59*, *UBE2C*	*SMAD2*, *RUNX2*, *ABTB1*, *ST5*, *CEBPB*, *SETDB1*, *CEBPG*
BRCA			*JAK2*, *NFKBIA*, *TBP*, *RXRA*, *VAV1*, *HES5*, *NFKBIB*
COAD (READ)			*EEF1A1*, *FOXG1*, *GADD45G*, *MAPK9*, *MYOC*, *SMAD2*
HNSC			*DUSP16*, *KRT8*, *RAF1*, *MED1*, *PPARG*, *YWHAB*, *FABP1*
LIHC			—
LUAD			*DOK1*, *FUT4*, *INSR*, *ITGB2*, *SHC1*, *PTPRC*, *KHDRBS1*
LUSC			*BRCA1*, *ETS2*, *HIF1A*, *JUN*, *LMO4*, *PIAS3*, *RBBP7*
KICH		—	—
KIRC			*AR*, *HGS*, *RUNX1*, *BCL3*, *BRCA1*, *STAT2*, *ITGA8*
KIRP			—
PRAD			—
THCA			—

Table 2. Gene expression biomarkers identified by different studies.

this with the remaining sample, was applied to each UGCB set. We first compared our 14 UGCBs (named Wx-14-UGCB) with the UGCBs (named Peng-14-UGCB) identified by Peng *et al.*[12] across seven cancer subtypes (BLCA, BRCA, COAD, HNSC, LIHC, LUAD, and LUSC). The Wx-14-UGCB set, which was identified by a neural network-based feature selection algorithm Wx, showed higher classification accuracy than Peng-14-UGCB for five out of seven different cancer types (Table 3).

The differentially expressed gene (DEG) analysis is typically used as a standard procedure when comparing transcriptomes (whole genes) between two (or more) conditions[18]. Therefore, we compared the Wx-14-UGCB with the top 14 DEGs (named DEG-14-UGCB; sorted into ascending order of adjusted *p* value) identified using a popular DEG analysis method called edgeR[7]. Similar to the above comparison, Wx-14-UGCB showed higher classification accuracy than DEG-14-UGCB for 7 different cancer types (Table 3). The area under the curve (AUC) values of BRCA, LUAD, and LUSC were 0.9944, 0.9943, and 0.9936, respectively (Fig. 2), showing excellent classification performance of the Wx-14-UGCB set.

We further evaluated the identified UGCB (by the Wx algorithm) by comparing those reported by Martinez-Ledesma *et al.*[13] (MartinezL-7-UGCB). Wx-7-UGCB showed higher accuracy than MartinezL-7-UGCB for five out of six cancer types (Table 3). We repeated the evaluation with a different algorithm called support-vector machine (SVM) (Table S2). The result showed that XGBoost achieved higher classification accuracy compared to SVM with the same UGCB set. The overall trend of classification accuracy is almost the same in XGBoost and SVM. Overall, the Wx-14-UGCB set, which was identified using the neural network-based feature selection algorithm Wx, was comparable to or outperformed previously reported universal gene expression biomarkers in terms of classification accuracy, highlighting the Wx algorithm's importance.

Putative role of the top 100 UGCBs. As shown in Fig. 1, approximately the top 100 UGCBs (Wx-100-UGCB) reached a plateau with the highest classification accuracy. We wondered how many genes identified by the Wx algorithm coincided with DEGs identified using edgeR. Intriguingly, less than 35% of genes overlapped (Fig. 3). For example, a comparison of the top 500 biomarker candidate genes identified by both algorithms showed that only 45 genes (9.0%) were common. In the case of top 2,000 genes, only 379 genes (19.0%) overlapped. Thus, there was substantial discrepancy between the algorithms with same gene expression data. Next, we performed gene ontology (GO) and network analysis to investigate the putative function of top 50 UGCBs using Metascape[19]. Genes involved in the Fc gamma receptor dependent phagocytosis, antigen processing and presentation, and regulation of apoptotic signaling pathway were significantly altered (Fig. S1), suggesting that the deregulation of these pathways might be a critical factor in the onset or progression of most cancers. Further investigations of these genes are warranted.

# of UGCBs	14			7	
Type	Wx	Peng's	edgeR	Wx	Martinez-Ledesma's
BLCA	95.79	97.20	94.86	95.79	96.26
BRCA	98.19	96.38	91.78	97.20	91.45
COAD	94.51	87.20	98.78	92.68	—
HNSC	97.17	92.23	94.35	95.05	92.57
KICH	95.65	—	100.00	97.83	—
KIRC	99.67	—	99.34	98.68	90.09
KIRP	99.38	—	99.38	100.00	—
LIHC	90.57	94.81	87.74	88.21	—
LUAD	97.92	97.58	98.96	97.92	90.27
LUSC	98.19	96.75	99.28	97.83	94.56
PRAD	93.45	—	92.36	90.55	—
THCA	95.80	—	90.21	95.45	
Total	96.72	94.59	94.81	95.74	92.20

Table 3. Classification accuracy comparison (%).

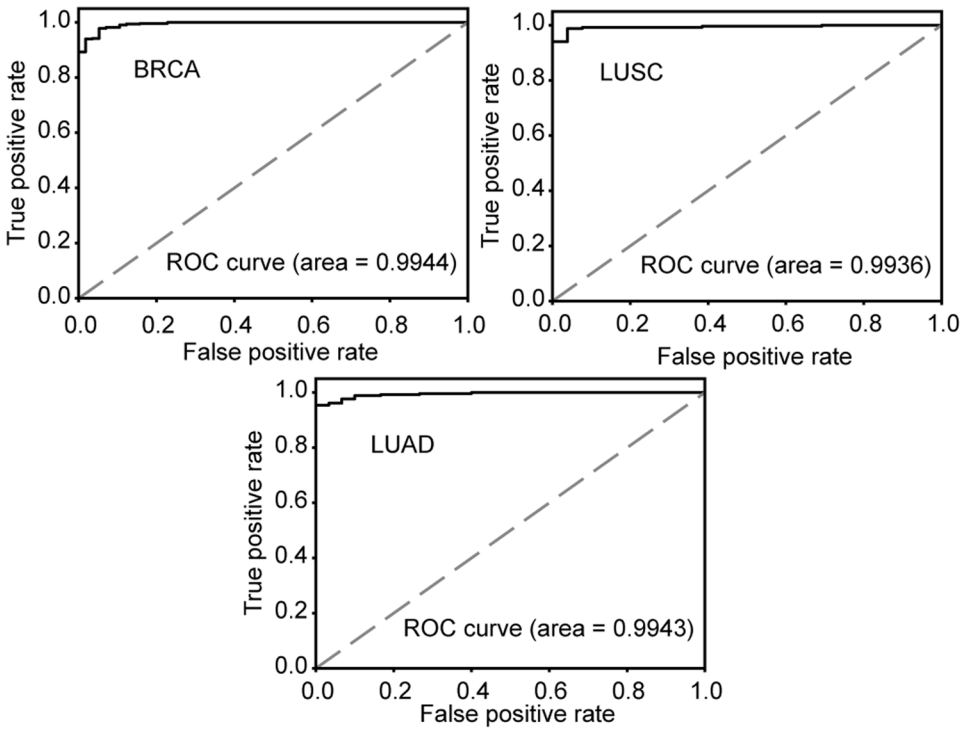

Figure 2. Performance of the Wx-14-UGCB on BRCA, LUAD, and LUSC RNA-seq data. AUC values are listed. ROC, receiver operating characteristic.

Additional validation of the identified UGCBs. Our comparison revealed that UGCBs identified by the Wx algorithm were comparable to or outperformed the UGCBs identified by different methods. We further validated the performance by evaluating the classification accuracy of Wx-14-UGCB and Peng-14-UGCB with cancer and normal RNA-seq data from three independent cancer studies including a melanoma cohort (GSE72056) that had not been included in the 12 types of TCGA cancer cohort[14,15] (Table 4). We calculated the classification accuracy by dividing the samples in a given cohort into the training set (2,888 samples, 64%), validation set (723 samples, 16%), and test set (902 samples, 20%). Then, the training set was used to train a model using a neural network (NN) algorithm and the validation set was used to assess how well the model had been trained. Finally, the test set was used to calculate the classification accuracy with the trained model. The comparison revealed that Wx-14-UGCB classified malignant and non-malignant melanoma single cells better than Peng-14-UGCB (Table 4). With the expression levels of the genes in the Wx-14-UGCB set, 818 out of the 902 test samples were correctly classified, whereas 633 out of 902 test samples were correctly classified using the Peng-14-UGCB set. For the lung adenocarcinoma data set (GSE40419), Wx-14-UGCB showed 80.00% classification accuracy when classifying lung cancer and adjacent normal cells, while Peng-14-UGCB showed 56.87% classification accuracy. For the 5,578 head and neck squamous cell carcinoma single cells (2,215 cancer cells and 3,363 non-cancer cells)

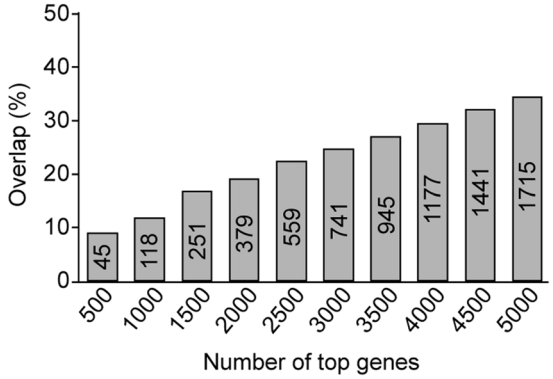

Figure 3. Comparison of genes identified by Wx and edgeR. The x-axis indicates the number of top genes used for the comparison and the y-axis represents the percentage of overlap between the gene sets.

GSE id	Cancer type	Wx-14-UGCB	Peng-14-UGCB
GSE72056	Melanoma	90.71	70.22
GSE40419	Lung adenocarcinoma	80.00	56.87
GSE103322	Head and neck squamous cell carcinoma (primary tumors and lymph node metastases; single-cell transcriptomes	81.10	68.28

Table 4. The classification accuracy of the UGCBs identified by different methods.

(GSE103322), Wx-14-UGCB showed 81.10% classification accuracy when classifying cancer and non-cancer cells, while Peng-14-UGCB showed 68.28% classification accuracy. In summary, the top 14 genes (Wx-14-UGCB) identified by the Wx algorithm could potentially be used as novel gene expression biomarkers for the detection of various types of cancers, although its use might be limited by clinical difficulties associated with RNA-based applications. Further experimental investigations are required to validate the Wx-14-UGCB.

Discussion

The next-generation sequencing (NGS) technique has opened up a new era in investigating genes and genomes by generating genome-wide molecular maps including the genome, transcriptome, and epigenome. Demand for NGS in many research fields has been growing rapidly since NGS can be used as a new kind of microscope, transforming information of entire molecules into numeric values[20]. However, this approach has given rise to another difficult problem; selecting appropriate genes (or loci) for directing the next step of a given study. For example, in the case of the human genome, selecting reasonable genes (features) from a list of expression levels over approximately 60,000 genes (or up to 190,000 transcripts) has become a major bottleneck. Many researchers have selected genes from a list of differentially expressed genes that is (DEGs) typically identified by a DEG identification algorithm with an adjusted *p* value of 0.05 (or less) for multiple tests. However, as the number of samples increases, the number of DEGs tends to increase, up to several thousand genes. Therefore, there is a demand for a method that automatically recommends the ideal gene set for biomarker candidates.

In this study, we have developed a neural network-based feature selection algorithm called Wx. The Wx algorithm provides a discriminative index (*DI*) score for each gene. The higher the *DI* score, the greater its influence on the classification of the given two groups of samples. Thus, when selecting genes for biomarkers, researchers can select the highest genes sorted (in descending order) by the *DI* score, and this can guarantee the highest classification accuracy, as shown in this study.

The 14 gene signatures (Wx-14-UGCB) identified by the Wx algorithm included the housekeeping gene *GAPDH*, which has been used in many studies as a control (or reference) gene (Table 2). Recently, several concerns about using the *GAPDH* gene as a housekeeping gene has been reported[21–25]. Our result also indicated that the *GAPDH* gene was one of the highest *DI*-score genes, and this gene should therefore be used with caution as a control gene in gene expression experiments such as qRT-PCR. Interestingly, another well-known housekeeping gene *ACTB* was ranked 14 out of 20,501 genes (Table S1), suggesting that both *GAPDH* and *ACTB* genes might be unsuitable housekeeping genes for gene expression experiments, particularly in cancer studies. The expression levels of the *GAPDH* and *ACTB* genes and the genes in the Wx-14-UGCB set in various cancer types also confirmed the variable expression levels of those genes between cancer and normal samples (http://firebrowse.org). Further investigations of the remaining genes such as *FN1*, *EEF1A1*, *COL1A1*, *SFTPB*, *SFTPC*, and *ATP1A1* will shed light on the identification of novel biomarker genes for a pan-cancer cohort.

One of the disadvantages of artificial neural network-based approaches when applied to biomedical data is that a large number of samples are needed to achieve good classification or regression performance. We observed relatively lower classification accuracy (Table S3) when the Wx algorithm was applied to a non-cancer transcriptomic data set (GSE105127), which contains normalized expression levels of 65,671 transcripts performed in the pericentral (n = 19), intermediate (n = 19), and periportal (n = 19) regions of the human liver isolated by

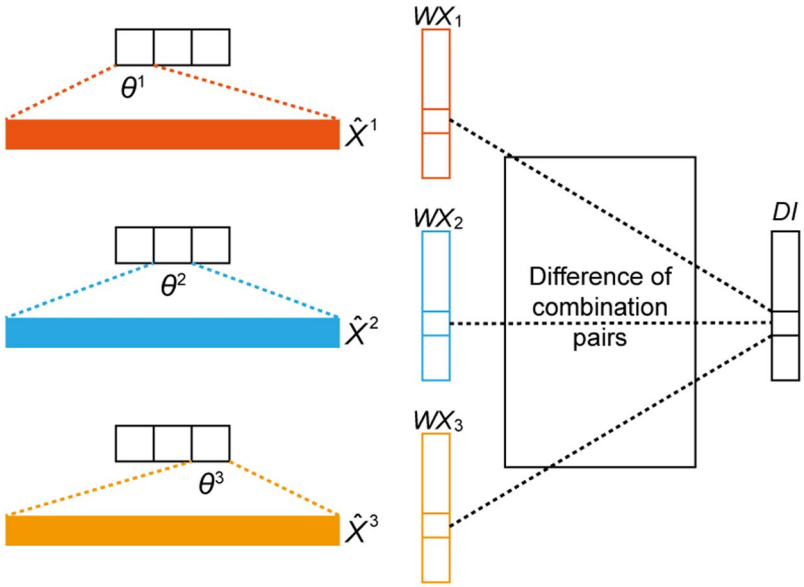

Figure 4. Discriminative index (*DI*) vector construction for $K = 3$, where $\boldsymbol{\theta^k}$ represents the parameter related to the k-th softmax output value, $\hat{\boldsymbol{X}}^{\boldsymbol{k}}$ is the averaged vector from all data samples with label $\boldsymbol{k}$, and $\boldsymbol{WX_k}$ the result of a multiplication between $\boldsymbol{\theta^k}$ and $\hat{\boldsymbol{X}}^{\boldsymbol{k}}$.

laser-captured microdissection[26]. In addition, selected features from the same data set vary depending on algorithms. In our comparison, there were no overlaps between the top 14 genes identified by Wx or Peng's (Table S2). This kind of inconsistency is caused mainly by the algorithmic difference, as reported in several differentially expressed gene analysis studies[18,27,28]. Thus, it is difficult to establish which algorithm is better by comparison without experimental verification. Therefore, the usefulness of the 14 genes (Wx-14-UGCB) for cancer biomarkers should be validated with extensive experimental evidence in the near future.

In summary, the Wx algorithm developed in this study estimates the classification power of genes in a given gene expression data set using the discriminative index (*DI*) score algorithm. Researchers can intuitively select gene-expression biomarker candidates from the *DI* scored gene list. Further experimental validation will be necessary to prove the Wx algorithm's usefulness.

Methods

Gene expression data sets used in this study. Gene expression data (mRNASeq) of 12 different cancer types from the cancer genome atlas (TCGA) were obtained using TCGA-Assembler 2[29]. Data generated by Illumina HiSeq instrument (labeled as illuminahiseq_rnaseqv2-RSEM_genes_normalized) were used in this study. Each sample contains normalized expression levels of 20,502 genes (features). A description of the TCGA data can be found in Table 1. The following independent RNA-seq data were used for validation; GSE40419[16] consists of normalized expression levels of 36,741 genes performed in 164 samples (87 lung cancer and 77 adjacent normal tissues) and GSE105127 includes normalized expression levels of 65,671 transcripts performed in the pericentral (n = 19), intermediate (n = 19), and periportal (n = 19) regions of human liver isolated by laser-captured microdissection[26]. These data sets were processed using Octopus-toolkit[30]. GSE72056[15] contains normalized expression levels of 23,686 genes performed in 1,257 malignant and 3,256 benign samples. GSE103322 contains normalized expression levels of 23,686 genes, performed in 5,578 head and neck squamous cell carcinoma single cells (2,215 cancer cells and 3,363 non-cancer cells)[17]. Gene expression tables for these data sets were downloaded from the gene expression omnibus website (https://www.ncbi.nlm.nih.gov/geo/).

Training and validation data sets. The gene expression data of 12 different cancer types includes 6,226 samples in total (5,609 cancer and 617 normal samples). The number of cancer and normal samples differs for each type of data, as shown in Table 1. In general, the number of cancer samples was much larger than the normal samples. Therefore, if we randomly divide samples into two groups (training and validation sets) without considering the ratio of cancer and normal samples, both groups will contain different ratios of cancer and normal samples. This could be problematic when training a model using a neural network. We avoided this imbalance by randomly dividing samples in each cancer set in half while maintaining the ratio of cancer and normal samples. One set was used for feature selection and the other was used for validation.

Model definition. The proposed feature selection method was based on softmax regression[31], which utilizes a simple one-layer neural network regression model in which the dependent variable is categorical. This model was applied to the feature selection set X^f and the validation set X^v; the details of each process are described below.

Let X be N number of gene expressions for tumor or normal samples, then it can be formally expressed as $X=\{X_1, X_2, \ldots, X_N\}$. Each X_i has J number of features ($X_i \in \Re^J$), each of which conveys information regarding the total expression amount of the corresponding gene. The output value $Y \in \Re^K$ is a one hot vector that consists of K numbers depending on how many classes it represents. In formal notation, the vector Y can be expressed as $Y=[y_1, y_2, \ldots, y_K]$. For example, if the problem is to classify tumor samples out of normal samples, the i-th input data with gene expression becomes $X_i=[x_{i1}, x_{i2}, \ldots, x_{i,J}]$, and the output becomes y_i. If the i-th data is from a normal sample, then $y_i=[1, 0]$, otherwise $y_i=[0, 1]$.

Softmax regression includes model parameters $\Theta=[\theta_1, \theta_2, \ldots, \theta_K]$ that are learned from the training data. With these parameters, the output Y_i can be expressed as Eq. 1 along with input X_i.

$$
\begin{aligned}
Y_{i;\Theta}(X_i) &= \begin{bmatrix} P(y=[1,\ 0,\ \ldots,\ 0]|X_i;\Theta) \\ P(y=[0,\ 1,\ 0,\ \ldots,\ 0]|X_i;\Theta) \\ \vdots \\ P(y=[0,\ 0,\ \ldots,\ 0,\ 1]|X_i;\Theta) \end{bmatrix} \\
&= \frac{1}{\sum_{k=1}^{K}\exp(\theta_k^{\mathrm{T}}X_i)} \begin{bmatrix} \exp(\theta_1^{\mathrm{T}}X_i) \\ \exp(\theta_2^{\mathrm{T}}X_i) \\ \vdots \\ \exp(\theta_K^{\mathrm{T}}X_i) \end{bmatrix}
\end{aligned} \tag{1}
$$

Once the parameters have been learned, the most informative genes for cancer sample classification can be chosen using the Discriminative Index algorithm (Algorithm 1). This is described in more detail in Section 2.4.

Neural network-based feature selection algorithm: Wx. The softmax regression parameters, Θ are trained using the subset of the whole dataset for feature selection, X^f and Y^f (for simplicity, we refer to these without their superscripts in this subsection). These parameters and the subset of the dataset serve as the input of the feature selection algorithm, which is based on the Discriminative Index (DI). This algorithm (Algorithm 1) returns c number of important features (genes) that is sufficient to successfully perform a given task:

Algorithm 1. Get top c genes using DI.

input: X, Θ, c
output: c number of gene names

Let X^k *be the input vector with class label k;*
for $k \leftarrow 1$ **to** K **do**
  $\hat{X}^k \leftarrow average(X^k)$;
  $WX_k \leftarrow \theta_K^{\mathrm{T}}\hat{X}^k$;
End
for $j \leftarrow 1$ **to** J **do**
  $DI_j \leftarrow 0$;
  foreach *combination pair (a, b) in {1,2, …,K}* **do**
    $DI_j \leftarrow DI_j + |WX_a(j) - WX_b(j)|$
  End
end
$DI_{sort} \leftarrow Sorted\ DI$ *in descending order*;
Return top c gene names in DI_{sort};

- Classify X into K classes according to their corresponding Y, which is denoted as X^1, X^2, …, X^K (Fig. 4).
- For each X^k, take the average for all instances to form an average vector, $\hat{X}^k \in \Re^J$
- Calculate the inner product between the parameter related to the k-th softmax output value, θ^k and the average vector, $\hat{X}^k$, which is assigned to $W\hat{X}^k$.
- Calculate the DI for feature (gene) j. This step considers all possible combination pairs of K classes; an example with $K=3$ is illustrated in Fig. 4. The DI calculation of index j can be done with ${}_3C_2=3$ number of absolute value additions between different pairs.

- After the iteration (lines 6–11 in Algorithm 1), the resulting *DI* is a vector of size *J*. Each element in this vector is sorted to form the sorted index DI_{sort}.
- The top *c* features (genes) are selected based on top *c* indices in DI_{sort}.

The Wx algorithm is a fast algorithm that ranks about 20000 features in 15 seconds with a typical computer in case of 5000 samples (Table S4). The stand-alone and web versions of the Wx algorithm are available at https://github.com/deargen/DearWXpub and https://wx.deargendev.me/, respectively.

Evaluation of the classification performance of the selected genes. The classification performance of the selected features (genes) was evaluated with a validation dataset using a scalable end-to-end tree boosting system called XGBoost[32]; the validation set was a subset of the entire dataset. When training the classifier with this subset, only selected features (genes) were fed into the classifier as an input. In formal notation, these new inputs can be represented as $X^{v_{reduced}} \in \Re^{N \times c}$.

Leave-one-out cross validation. Leave-one-out cross validation (LOOCV) was used to assess whether a given set of UCBs could be used to distinguish between normal and cancer samples.

Gene ontology and network analysis. Metascape[19] was used to infer key functions of top 50 genes ranked by the *DI* score. MCODE algorithm[33] was then applied to identify key networks.

References

1. Perou, C. M. *et al*. Molecular portraits of human breast tumours. *Nature* **406**, 747–752, https://doi.org/10.1038/35021093 (2000).
2. Cancer Genome Atlas, N. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70, https://doi.org/10.1038/nature11412 (2012).
3. Ptitsyn, A., Hulver, M., Cefalu, W., York, D. & Smith, S. R. Unsupervised clustering of gene expression data points at hypoxia as possible trigger for metabolic syndrome. *BMC Genomics* **7**, 318, https://doi.org/10.1186/1471-2164-7-318 (2006).
4. Dennis, G. Jr. *et al*. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* **4**, P3 (2003).
5. Zou, H. & Hastie, T. Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B* **67**, 301–320 (2005).
6. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq 2. *Genome Biol* **15**, 550, https://doi.org/10.1186/s13059-014-0550-8 (2014).
7. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140, https://doi.org/10.1093/bioinformatics/btp616 (2010).
8. Conesa, A. *et al*. A survey of best practices for RNA-seq data analysis. *Genome Biol* **17**, 13, https://doi.org/10.1186/s13059-016-0881-8 (2016).
9. Rohart, F., Gautier, B., Singh, A. & Le Cao, K. A. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput Biol* **13**, e1005752, https://doi.org/10.1371/journal.pcbi.1005752 (2017).
10. Perez-Riverol, Y., Kuhn, M., Vizcaino, J. A., Hitz, M. P. & Audain, E. Accurate and fast feature selection workflow for high-dimensional omics data. *PLoS One* **12**, e0189875, https://doi.org/10.1371/journal.pone.0189875 (2017).
11. Cancer Genome Atlas Research, N. *et al*. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45**, 1113–1120, https://doi.org/10.1038/ng.2764 (2013).
12. Peng, L. *et al*. Large-scale RNA-Seq Transcriptome Analysis of 4043 Cancers and 548 Normal Tissue Controls across 12 TCGA Cancer Types. *Sci Rep* **5**, 13413, https://doi.org/10.1038/srep13413 (2015).
13. Martinez-Ledesma, E., Verhaak, R. G. & Trevino, V. Identification of a multi-cancer gene expression biomarker for cancer clinical outcomes using a network-based algorithm. *Sci Rep* **5**, 11966, https://doi.org/10.1038/srep11966 (2015).
14. Yu, K. *et al*. A precisely regulated gene expression cassette potently modulates metastasis and survival in multiple solid cancers. *PLoS Genet* **4**, e1000129, https://doi.org/10.1371/journal.pgen.1000129 (2008).
15. Tirosh, I. *et al*. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196, https://doi.org/10.1126/science.aad0501 (2016).
16. Seo, J. S. *et al*. The transcriptional landscape and mutational profile of lung adenocarcinoma. *Genome Res* **22**, 2109–2119, https://doi.org/10.1101/gr.145144.112 (2012).
17. Puram, S. V. *et al*. Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck. *Cancer. Cell* **171**, 1611–1624, https://doi.org/10.1016/j.cell.2017.10.044 (2017).
18. Finotello, F. & Di Camillo, B. Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. *Brief Funct Genomics* **14**, 130–142, https://doi.org/10.1093/bfgp/elu035 (2015).
19. Tripathi, S. *et al*. Meta- and Orthogonal Integration of Influenza "OMICs" Data Defines a Role for UBR4 in Virus Budding. *Cell Host Microbe* **18**, 723–735, https://doi.org/10.1016/j.chom.2015.11.002 (2015).
20. Shendure, J. *et al*. DNA sequencing at 40: past, present and future. *Nature* **550**, 345–353, https://doi.org/10.1038/nature24286 (2017).
21. Glare, E. M., Divjak, M., Bailey, M. J. & Walters, E. H. beta-Actin and GAPDH housekeeping gene expression in asthmatic airways is variable and not suitable for normalising mRNA levels. *Thorax* **57**, 765–770 (2002).
22. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends Genet* **29**, 569–574, https://doi.org/10.1016/j.tig.2013.05.010 (2013).
23. Barber, R. D., Harmer, D. W., Coleman, R. A. & Clark, B. J. GAPDH as a housekeeping gene: analysis of GAPDH mRNA expression in a panel of 72 human tissues. *Physiol Genomics* **21**, 389–395, https://doi.org/10.1152/physiolgenomics.00025.2005 (2005).
24. Sikand, K., Singh, J., Ebron, J. S. & Shukla, G. C. Housekeeping gene selection advisory: glyceraldehyde-3-phosphate dehydrogenase (GAPDH) and beta-actin are targets of miR-644a. *PLoS One* **7**, e47510, https://doi.org/10.1371/journal.pone.0047510 (2012).
25. Caradec, J., Sirab, N., Revaud, D., Keumeugni, C. & Loric, S. Is GAPDH a relevant housekeeping gene for normalisation in colorectal cancer experiments? *Br J Cancer* **103**, 1475–1476, https://doi.org/10.1038/sj.bjc.6605851 (2010).
26. Brosch, M. *et al*. Epigenomic map of human liver reveals principles of zonated morphogenic and metabolic control. *Nat Commun* **9**, 4150, https://doi.org/10.1038/s41467-018-06611-5 (2018).
27. Costa-Silva, J., Domingues, D. & Lopes, F. M. RNA-Seq differential expression analysis: An extended review and a software tool. *PLoS One* **12**, e0190152, https://doi.org/10.1371/journal.pone.0190152 (2017).
28. Wang, T., Li, B., Nelson, C. E. & Nabavi, S. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinformatics* **20**, 40, https://doi.org/10.1186/s12859-019-2599-6 (2019).
29. Wei, L. *et al*. TCGA-Assembler 2: Software Pipeline for Retrieval and Processing of TCGA/CPTAC Data. *Bioinformatics*, https://doi.org/10.1093/bioinformatics/btx812 (2017).

30. Kim, T., Seo, H. D., Hennighausen, L., Lee, D. & Kang, K. Octopus-toolkit: a workflow to automate mining of public epigenomic and transcriptomic next-generation sequencing data. *Nucleic Acids Res* **46**, e53, https://doi.org/10.1093/nar/gky083 (2018).
31. Peduzzi, P., Concato, J., Kemper, E., Holford, T. R. & Feinstein, A. R. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol* **49**, 1373–1379 (1996).
32. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, https://doi.org/10.1145/2939672.2939785 (2016).
33. Bader, G. D. & Hogue, C. W. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* **4**, 2 (2003).

Acknowledgements

We are grateful to Drs. Young-Ho Ahn and Ji Hyung Hong who moderated this paper. This study was supported by a grant from the National R&D Program for Cancer Control, Ministry of Health & Welfare, Republic of Korea (1720100).

Author Contributions

Concept and design: K.K.*, S.P., B.S.; Development of the algorithm: S.P., B.S., W.S.S.; Analysis and interpretation: K.K.*, S.P., B.S., K.K., Y.C.; Drafting the manuscript: all authors.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-47016-8.

Competing Interests: Dr. Kang (K.K*) is one of the co-founders of, and a shareholder in, Deargen Inc.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.